Fitness tracking reveals task-specific associations between memory, mental health, and exercise

Jeremy R. Manning[1,⋆], Gina M. Notaro[1,2], Esme Chen[1], and Paxton C. Fitzpatrick[1]

[1]Dartmouth College, Hanover, NH

[2]Lockheed Martin, Bethesda, MD

⋆Address correspondence to jeremy.r.manning@dartmouth.edu

October 6, 2021

Abstract

Physical exercise can benefit both physical and mental well-being. Different forms of exercise (i.e., aerobic versus anaerobic; running versus walking versus swimming versus yoga; high-intensity interval training versus endurance workouts; etc.) impact physical fitness in different ways. For example, running may substantially impact leg and heart strength but only moderately impact arm strength. We hypothesized that the mental benefits of exercise might be similarly differentiated. We focused specifically on how different forms of exercise might related to different aspects of memory and mental health. To test our hypothesis, we collected nearly a century's worth of fitness data (in aggregate). We then asked participants to fill out surveys asking them to self-report on different aspects of their mental health. We also asked participants to engage in a battery of memory tasks that tested their short and long term episodic, semantic, and spatial memory. We found that participants with similar exercise habits and fitness profiles tended to also exhibit similar mental health and task performance profiles.

Introduction

Engaging in physical activity (exercise) can improve our physical fitness by increasing muscle strength (Crane et al., 2013; Knuttgen, 2007; Lindh, 1979; Rogers and Evans, 1993), increasing bone density (Bassey and Ramsdale, 1994; Chilibeck et al., 2012; Layne and Nelson, 1999), increasing cardiovascular performance (Maiorana et al., 2000; Pollock et al., 2000), increasing lung capacity (Lazovic-Popovic et al., 2016) (although see Roman et al., 2016), increasing endurance (Wilmore and Knuttgen, 2003), and more. Exercise can also improve mental health (Basso and Suzuki, 2017; Callaghan, 2004; Deslandes et al., 2009; Mikkelsen et al., 2017; Paluska and Schwenk, 2000; Raglin, 1990; Taylor et al., 1985) and cognitive performance (Basso and Suzuki, 2017; Brisswalter et al., 2002; Chang et al., 2012; Etnier et al., 2006).

The physical benefits of exercise can be explained by stress-responses of the affected body tissues. For example, skeletal muscles that are taxed during exercise exhibit stress responses (Morton et al., 2009) that can in turn affect their growth or atrophy (Schiaffino et al., 2013). By comparison, the benefits of exercise on mental health are less direct. For example, one hypothesis is that exercise leads to specific physiological changes, such as increased aminergic synaptic transmission and endorphin release, which in turn act on neurotransmitters in the brain (Paluska and Schwenk, 2000).

Speculatively, if different exercise regimens lead to different neurophysiological responses, one might be able to map out a spectrum of signalling and transduction pathways that are impacted by a given type, duration, and intensity of exercise in each brain region. For example, prior work has shown that exercise increases acetylcholine levels, starting in the vicinity of the exercised muscles (Shoemaker et al., 1997). Acetylcholine is thought to play an important role in memory formation (Palacios-Filardo et al., 2021, e.g., by modulating specific synaptic inputs from entorhinal cortex to the hippocampus, albeit in rodents;). Given the central role of these medial temporal lobe structures play in memory, changes in acetylcholine might lead to specific changes in memory formation and retrieval.

In the present study, we hypothesize that (a) different exercise regimens will have different,

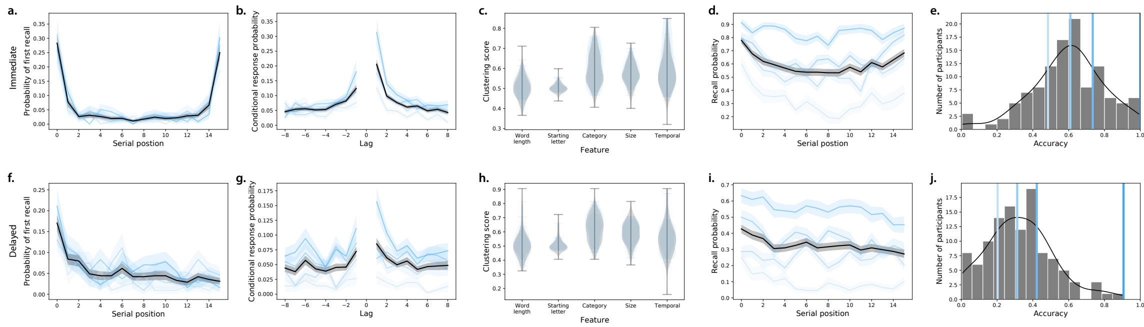

Figure 1: **Free recall behavioral results.**

quantifiable impacts on cognitive performance and mental health, and that (b) these impacts will be consistant across individuals. To this end, we collected a year of fitness tracking data from each of 113 participants. We then asked each participant to fill out a brief survey in which they self-evaluated several aspects of their mental health. Finally, we ran each participant through a battery of memory tasks, which we used to evaluate their memory performance along several dimensions. We examined the data for potential associations between memory, mental health, and exercise.

Results

Before testing our main hypothesis, we first examined the behavioral data from each memory task. We expected that the general trends and tendancies in the behavioral data would follow previously reported behaviors from similar tasks that had been utilized in prior work. We were also interested in characterizing the variability in task performance across participants. For example, if all participants exhibited near-identical behaviors or performance on a given task, we would be unable to identify how memory performance varied with mental health or exercise.

- characterizing behaviors (color by quartile and continue the color scheme in later figures– hue reflects task, shading reflects performance. white outline means immediate, black outline means delayed)

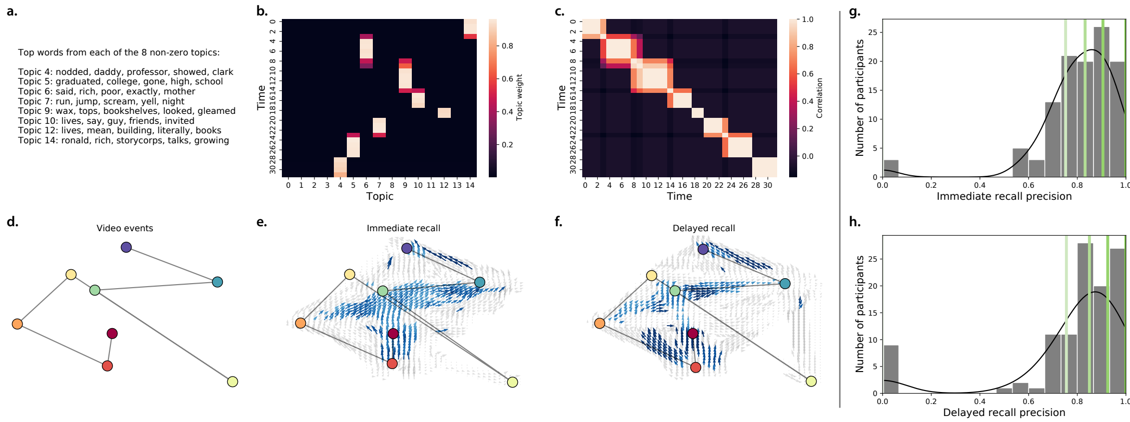

Figure 2: **Naturalistic recall behavioral results.**

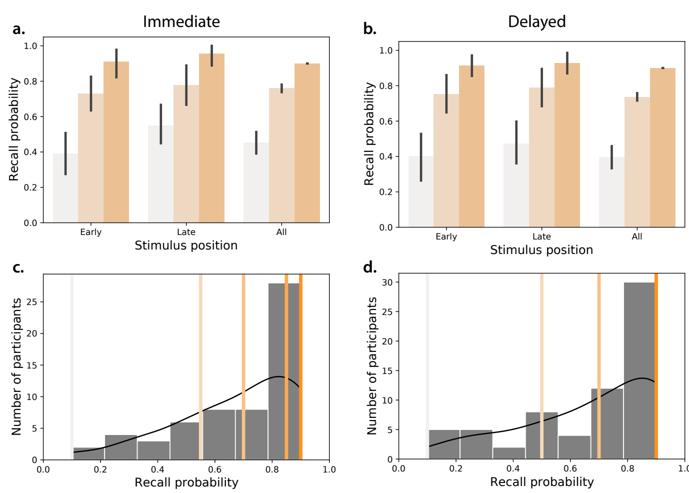

Figure 3: **Foreign language vocabulary learning behavioral results.**

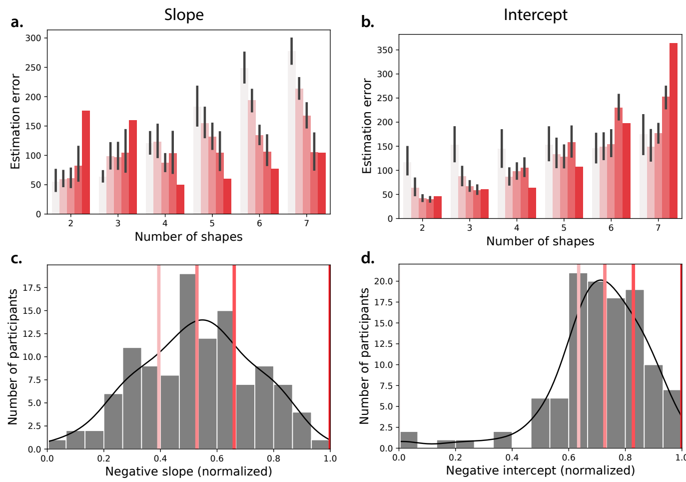

Figure 4: **Spatial learning behavioral results.**

  - Free recall (immediate + delayed): pfr, lag-CRP, spc (color: recall performance). Figure 1.
  - Naturalistic recall (immediate + delayed): reproduce a version of the sherlock movie/recall trajectories (color: mean precision). Figure 2.
  - Foreign language flashcards (immediate + delayed): p(correct) histogram (color: p(correct)). Figure 3.
  - Spatial learning: mean error by number of shapes (color: intercept and slope of line fit to errors as a function of the number of shapes). Figure 4.

- Fitness info (break down by task performance, potentially separately for each task); also separate out recent (raw) and recent versus baseline – color using same color scheme as behavior figure
  - activity (steps, zone minutes, floors/elevation)
  - resting heart rate
  - sleep
- exploratory analysis (correlations). Possibly make some sort of scatter matrix or pairplot– rows/columns: tasks. Diagonal: histogram of performance metric for that task. Above-

diagonal entries: compare performance across tasks (by subject). Below-diagonal entries: empty? density/2d histograms?

  - Memory-memory
  - fitness-fitness
  - survey-survey
  - (fitness + survey)-memory

- predictive analysis (regressions)
  - Predict memory performance on held-out task from other tasks
  - Predict memory performance on each task using fitness data
  - Predict memory performance on each task using survey data

- Reverse correlations: look at recent changes versus baseline trends (color using same scheme as behavior figures). Possibly
  - Fitness profile that predicts performance on each task (barplots + timelines)
  - Fitness profile for each survey demographic (barplots + timelines)
    * Select out mental health demographics (based on meds, stress levels)

Discussion

- summarize key findings
- correlation versus causation
- what can vs. can't we know? we can identify correlations, but not causal direction– e.g. we cannot know whether exercise *causes* mental changes versus whether people with particular neural profiles might tend to engage in particular exercise behaviors. that being said, we *can* separate out baseline tendencies (e.g., how people tend to exercise in general) versus recent changes (e.g., how they happened to have exercised prior to the experiment).

- related work (exercise/memory, exercise/mental health), what this study adds
- future direction: towards customized physical exercise recommendation engine for optimizing mental health and mental fitness

Methods

We ran an online experiment using the Amazon Mechanical Turk platform. We collected data about each participant's fitness and exercise habits, a variety of self-reported measures concerning their mental health, and about their performance on a battery of memory tasks. We mined the dataset for potential associations between memory, mental health, and exercise.

Experiment

Participants

We recruited experimental participants by posting our experiment as a Human Intelligence Task (HIT) on the Amazon Mechanical Turk platform. We limited participation to Mechanical Turk Workers who had been assigned a "Masters" designation on the platform, given to workers who score highly across several metrics on a large number of HITs, according to a proprietary algorithm managed by Amazon. We further limited our participant pool to participants who self-reported that they were fluent in English and regularly used a Fitbit fitness tracker device. A total of 160 workers accepted our HIT in order to participate in our experiment. Of these, we excluded all participants who failed to log into their Fitbit account (giving us access to their anonymized fitness tracking data), encountered technical issues (e.g., by accessing the HIT using an incompatible browser, device, or operating system), or who ended their participation prematurely, before completing the full study. In all, 113 participants remained that contributed usable data to the study.

For their participation, workers received a base payment of $5 per hour (computed in 15 minute increments, rounded up to the nearest 15 minutes), plus an additional performance-based bonus of up to $5. Our recruitment procedure and study protocol were approved by Dartmouth's

Committee for the Protection of Human Subjects.

Gender, age, and race. Of the 113 participants who contributed usable data, 77 reported their gender as female, 35 as male, and 1 chose not to report their gender. Participants ranged in age from 19–68 years old (25^{th} percentile: 28.25 years; 50^{th} percentile: 32 years; 75^{th} percentile: 38 years). Participants reported their race as White (90 participants), Black or African American (11 participants), Asian (7 participants), Other (4 participants), and American Indian or Alaska Native (3 participants). One participant opted not to report their race.

Languages. All participants reported that they were fluent in either 1 and 2 languages (25^{th} percentile: 1; 50^{th} percentile: 1; 75^{th} percentile: 1), and that they were "familiar" with between 1 and 11 languages (25^{th} percentile: 1; 50^{th} percentile: 2; 75^{th} percentile: 3).

Reported medical conditions and medications. Participants reported having and/or taking medications pertaining to the following medical conditions: anxiety or depression (4 participants), recent head injury (2 participants), high blood pressure (1 participant), bipolar (1 participant), hypothyroidism (1 participant), and other unspecified medications (1 participant). Participants reported their current and typical stress levels on a Likert scale as very relaxed (-2), a little relaxed (-1), neutral (0), a little stressed (1), or very stressed (2). The "current" stress level reflected participants' stress at the time they participated in the experiment. Their responses ranged from -2 to 2 (current stress: 25^{th} percentile: -2; 50^{th} percentile: -1; 75^{th} percentile: 1; typical stress: 25^{th} percentile: 0; 50^{th} percentile: 1; 75^{th} percentile: 1). Participants also reported their current level of alertness on a Likert scale as very sluggish (-2), a little sluggish (-1), neutral (0), a little alert (1), or very alert (2). Their responses ranged from -2 to 2 (25^{th} percentile: 0; 50^{th} percentile: 1; 75^{th} percentile: 2). Nearly all (111 out of 113) participants reported that they had normal color vision, and 15 participants reported uncorrected visual impairments (including dyslexia and uncorrected near- or far-sightedness).

Residence and level of education. Participants reported their residence as being located in the suburbs (36 participants), a large city (30 participants), a small city (23 participants), rural (14 participants), or a small town (10 participants). Participants reported their level of education as follows: College graduate (42 participants), Master's degree (23 participants), Some college (21 participants), High school graduate (9 participants), Associate's degree (8 participants), Other graduate or professional school (5 participants), Some graduate training (3 participants), or Doctorate (2 participants).

Reported water and coffee intake. Participants reported the number of cups of water and coffee they had consumed prior to accepting the HIT. Water consumption ranged from 0–6 cups (25$^{\text{th}}$ percentile: 1; 50$^{\text{th}}$ percentile: 3; 75$^{\text{th}}$ percentile: 4). Coffee consumption ranged from 0–4 cups (25$^{\text{th}}$ percentile: 0; 50$^{\text{th}}$ percentile: 1; 75$^{\text{th}}$ percentile: 2).

Tasks

Upon accepting the HIT posted on Mechanical Turk, the worker was directed to read and fill out a screening and consent form, and to share access to their anonymized Fitbit data via their Fitbit account. After consenting to participant and successfully sharing their Fitbit data, participants filled out a survey and then engaged in a series of memory tasks (Fig. 5). All stimuli and code for running the full Mechanical Turk experiment may be found here.

Survey questions. We collected the following demographic information from each participant: their birth year, gender, highest (academic) degree achieved, race, language fluency, and language familiarity. We also collected information about participants' health and wellness, including about their vision, alertness, stress, sleep, coffee and water consumption, location of their residence, activity typically required for their job, and exercise habits.

Free recall (Fig. 5a). Participants studied a sequence of four word lists, each comprising 16 words. After studying each list, participants received an immediate memory test, whereby they were asked to type (one word at a time) any words they remembered from the just-studied list, in any order.

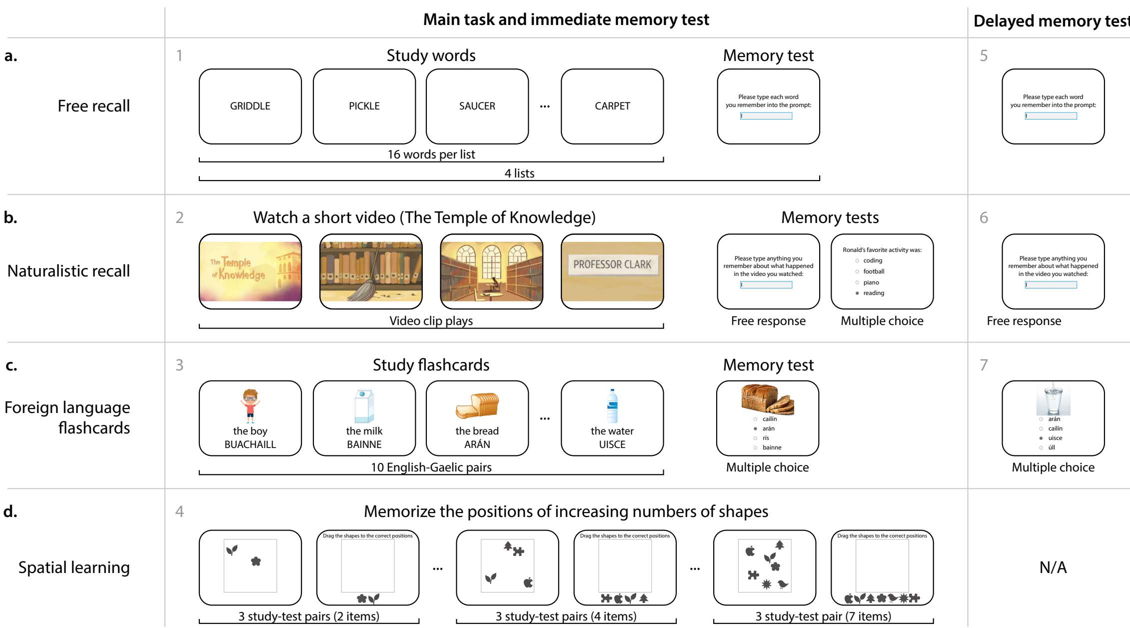

Figure 5: **Battery of memory tasks. a. Free recall.** Participants study 16 words (presented one at a time), followed by an immediate memory test where they type each word they remember from the just-studied list. In the delayed memory test, participants type any words they remember studying, from any list. **b. Naturalistic recall.** Participants watch a brief video, followed by two immediate memory tests. The first test asks participants to write out what happened in the video. The second test has participants answer a series of multiple choice questions about the conceptual content of the video. In the delayed memory test, participants (again) write out what happened in the video. **c. Foreign language flashcards.** Participants study a sequence of 10 English-Gaelic word pairs, each presented with an illustration of the given word. During an immediate memory test, participants perform a multiple choice test where they select the Gaelic word that corresponds to the given photograph. During the delayed memory test, participants perform a second multiple choice test, where they select the Gaelic word that corresponds to each of a new set of photographs. **d. Spatial learning.** In each trial, participants study a set of randomly positioned shapes. Next, the shapes' positions are altered, and participants are asked to drag the shapes back to their previous positions. **All panels.** The gray numbers denote the order in which participants experienced each task or test.

Words were presented for 2 s each, in black text on a white background, followed by a 2 s blank (white) screen. After the final 2 s pause, participants were given 90 s to type in as many words as they could remember, in any order. The memory test was constructed such that the participant could only see the text of the current word they were typing; when they pressed any non-letter key, the current word was submitted and the text box they were typing in was cleared. This was intended to prevent participants from retroactively editing their previous responses.

The word lists participants studied were drawn from the categorized lists reported in Ziman et al. (2018). Each participant was assigned four unique randomly chosen lists (in a randomized order), selected from a full set of 16 lists. Each chosen list was then randomly shuffled before presenting the words to the participants.

Participants also performed a final delayed memory test where they were given 180 s to type out any words they remembered from *any* of the 4 lists they had studied.

Recalled words within an edit distance of 2 (i.e., a Levenshtein Distance less than or equal to 2) of any word in the wordpool were "autocorrected" to their nearest match. We also manually corrected clear typos or misspellings by hand (e.g., we corrected "hippoptumas" to "hippopotamus", "zucinni" to "zucchini", and so on). Finally, we lemmatized each submitted word to match the plurality of the matching wordpool word (e.g., "bongo" was corrected to "bongos", and so on). After applying these corrections, any submitted words that matched words presented on the just-studied list were tagged as "correct" recalls, and any non-matching words were discarded as "errors." Because participants were not allowed to edit the text they entered, we chose not to analyze these putative "errors," since we could not distinguish typos from true misrememberings.

Naturalistic recall (Fig. 5b). Participants watched a 2.5 minute video clip entitled "The Temple of Knowledge." The video comprises an animated story told to StoryCorps by Ronald Clark, who was interviewed by his daughter, Jamilah Clark. The narrator (Ronald) discusses growing up living in an apartment over Washington Heights branch of the New York Public Library, where his father worked as a custodian during the 1940s.

After watching the video clip, participants were asked to type out anything they remembered about what happened in the video. They typed their responses into a text box, one sentence at a time. When the participant pressed the return key or typed any final punctuation mark (".", "!", or "?") the text currently entered into the box was "submitted" and added to their transcript, and the text box was cleared to prevent further editing of any already-submitted text. This was intended to prevent participants from retroactively editing their previous responses. Participants were given up to 10 minutes to enter their responses. After 4 minutes participants were given the option of ending the response period early, e.g., if they felt they had finished entering all of the information they remembered. Each participant's transcript was constructed from their submitted responses by combining the sentences into a single document and removing extraneous whitespace characters.

Following this 4–10 minute free response period, participants were given a series of 10 multiple choice questions about the conceptual content of the story. All participants received the same questions, in the same order.

Participants also performed a final delayed memory test, where they carried out the free response recall task a second time, near the end of the testing session. This resulted in a second transcript, for each participant.

Foreign language flashcards (Fig. 5c). Participants studied a series of 10 English-Gaelic word pairs in a randomized order. We selected the Gaelic language both for its relatively small number of native speakers and for its dissimilarity to other commonly spoken languages amongst Mechanical Turk Workers. We verified (via self report) that all of our participants were fluent in English and that they were neither fluent nor familiar with Gaelic.

Each word's "flashcard" comprised a cartoon depicting the given word, the English word or phrase in lowercase text (e.g., "the boy"), and the Gaelic word or phrase in uppercase text (e.g., "BUACHAILL"). Each flashcard was displayed for 4 s, followed by a 3 s interval (during which the screen was cleared) prior to the next flashcard presentation.

After studying all 10 flashcards, participants were given a multiple choice memory test where they were shown a series of novel photographs, each depicting one of the 10 words they had learned. They were asked to select which (of 4 unique options) Gaelic word went with the given picture. The 3 incorrect options were selected at random (with replacement across trials), and the order in which the choices appeared to the participant were also randomized. Each of the 10 words they had learned were tested exactly once.

Participants also performed a final delayed memory test, where they were given a second set of 10 questions (again, one per word they had studied). For this second set of questions participants were prompted with a new set of novel photographs, and new randomly chosen incorrect choices for each question. Each of the 10 original words they had learned were (again) tested exactly once during this final memory test.

Spatial learning (Fig. 5d). Participants performed a series of study-test trials where they memorized the onscreen spatial locations of two or more shapes. During the study phrase of each trial, a set of shapes appeared on the screen for 10 s, followed by 2 s of blank (white) screen. During the test phase of each trial, the same shapes appeared onscreen again, but this time they were vertically aligned and sorted horizontally in a random order. Participants were instructed to drag (using the mouse) each shape to its studied position, and then to click a button to indicate that the placements were complete.

In different study-test trials, participants learned the locations of different numbers of shapes (always drawn from the same pool of 7 unique shapes, where each shape appeared at most one time per trial). They first performed three trials where they learned the locations of 2 shapes; next three trials where they learned the locations of 3 shapes; and so on until their last three trials, where (during each trial) they learned the locations of 7 shapes. All told, each participant performed 18 study-test trials of this spatial learning task (3 trials for each of 2, 3, 4, 5, 6, and 7 shapes).

Fitness tracking using Fitbit devices

To gain access to our study, participants provided us with access to all data associated with their Fitbit account from the year (365 calendar days) up to and including the day they accepted the HIT. We filtered out all identifiable information (e.g., participant names, GPS coordinates, etc.) prior to importing their data.

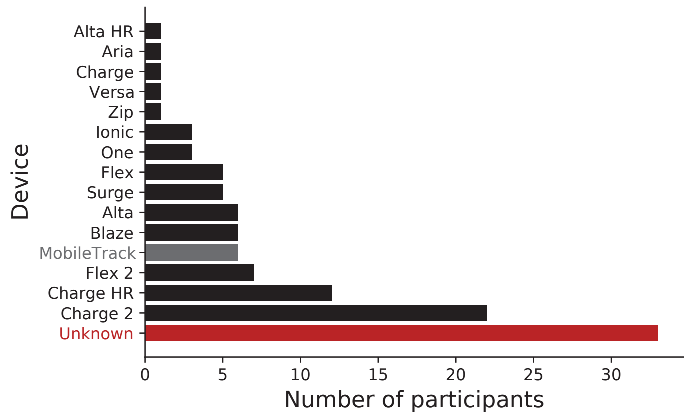

Figure 6: **Fitbit devices.** The bars indicate the numbers of participants whose fitness tracking data came from each model of Fitbit device. “MobileTrack” refers to participants who used smartphone accelerometer information to track their activity via the Fitbit smartphone app. “Unknown” denotes participants whose device information was not available from their available Fitbit data.

Collecting and processing Fitbit data

The fitness tracking data associated with participants' Fitbit accounts varied in scope and duration according to which device the participant owned (Fig. 6), how often the participant wore (and/or synced) their tracking device, and how long they had owned their device. For example, while all participants' devices supported basic activity metrics such as daily step counts, only a subset of the devices with heart rate monitoring capabilities provided information about workout intensity, resting heart rate, and other related measures.

Across all devices, we collected the following information: heart rate data, sleep tracking data, logged bodyweight measurements, logged nutrition measurements, Fitbit account and device settings, and activity metrics.

Heart rate. If available, we extracted all heart rate data collected by participants' Fitbit device(s) and associated with their Fitbit profile. Depending on the specific device model(s) and settings, this included second-by-second, minute-by-minute, daily summary, weekly summary, and/or monthly summary heart rate information. These summaries include information about participants' average heart rates, and the amount of time they were estimated to have spent in different “heart rate zones" (rest, out-of-range, fat burn, cardio, or peak, as defined by their Fitbit profile), as well as an estimate of the number of estimated calories burned while in each heart rate zone.

Sleep. If available, we extracted all sleep data collected by participants' Fitbit device(s). Depending on the specific device model(s) and settings, this included nightly estimates of the duration and quality of sleep, as well as the amount of time spent in each sleep stage (awake, REM, light, or deep).

Weight. If available, we extracted any weight-related information affiliated with participants' Fitbit accounts within 1 year prior to enrolling in our study. Depending on their specific device model(s) and settings, this included their weight, body mass index, and/or body fat percentage.

Nutrition. If available, we extracted any nutrition-related information affiliated with participants' Fitbit accounts within 1 year prior to enrolling in our study. Depending on their specific account settings and usage behaviors, this included a log of the specific foods they had eaten (and logged) over the past year, and the amount of water consumed each day.

Account and device settings. We extracted any settings associated with participants' Fitbit accounts to determine (a) which device(s) and model(s) are associated with their Fitbit account, (b) time(s) when their device(s) were last synced, and (c) battery level(s).

Activity metrics. If available, we extracted any activity-related information affiliated with participants' Fitbit accounts within 1 year prior to enrolling in our study. Depending on their specific device model(s) and settings, this included: daily step counts; daily amount of time spent in each activity level (sedentary, lightly active, fairly active, or very active, as defined by their account settings and preferences); daily number of floors climbed; daily elevation change; and daily total distance traveled.

Comparing recent versus baseline measurements.

We were interested in separating out potential associations between *absolute* fitness metrics and *relative* metrics. To this end, in addition to assessing potential raw (absolute) fitness metrics, we also defined a simple measure of recent changes in those metrics, relative to a baseline:

$$\Delta_{R,B} m = \frac{B \sum_{i=1}^{R} m(i)}{R \sum_{i=R+1}^{R+B} m(i)},$$

where $m(i)$ is the value of metric m from $i - 1$ days prior to testing (e.g., $m(1)$ represents the value of m on the day the participant accepted the HIT, and $m(10)$ represents the value of m 9 days prior to accepting the HIT. Unless otherwise noted, we set $R = 7$ and $B = 30$. In other words, to estimate recent changes in any metric m, we divided the average value of m taken over the prior week by the average value of m taken over the 30 days before that.

Exploratory correlation analyses

We used a bootstrap procedure to identify reliable correlations between different memory-related, fitness-related, and demographic-related variables. For each of $N = 1000$ iterations, we selected (with replacement) a sample of 113 participants to include. This yielded, for each iteration, a sampled "data matrix" with one row per sampled participant and one column for each measured variable. When participants were sampled multiple times in a given iteration, as was often the case, this matrix contained duplicate rows. We used a round-robin imputation procedure to estimate the values of any missing features (Buck, 1960). Next, we computed the Pearson's correlation between each pair of columns. This yielded, for each pair of columns, a distribution of N bootstrapped correlation coefficients. If fewer than 95% of the coefficients for a given pair of columns had the same sign, we excluded the pair from further analysis and considered the expected correlation between those columns to be undefined. If $\geq$ 95% of the coefficients for a given pair of columns had the same sign, we computed the expected correlation coefficient as:

$$\mathbb{E}_{i,j}\left[r\right] = \tanh\left(\frac{1}{N}\sum_{n=1}^{N}\tanh^{-1}\left(\mathrm{corr}\left(m(i)_n, m(j)_n\right)\right)\right),$$

where $m(x)_n$ represents column x of the bootstrapped data matrix for iteration n, tanh is the hyperbolic tangent, and $\tanh^{-1}$ is the inverse hyperbolic tangent.

Regression-based prediction analyses

Following our exploratory correlation analyses, we used an analogous bootstrap procedure to identify subsets of memory-related, fitness-related, and demographic-related variables that predicted (non-overlapping) subsets of other variables. For example, we tested whether a combination of fitness-related variables could predict a combination of memory-related variables, and so on.

We used the same bootstrap procedure described above (used in our exploratory correlation analyses) to generate $N = 1000$ bootstrapped data matrices whose rows reflected sampled participants and whose columns reflected different measured variables.

We grouped variables according to whether they were memory-related, fitness-related, or demographic-related. For each bootstrap iteration, we divided the rows of that iterations data matrix into training and test sets. The assignments of rows to these two sets was random, subject to the constraint that any duplicated rows in the data matrix (i.e., reflecting a single participant who had been sampled multiple times) was always assigned to either the training *or* the test set–i.e., duplicated rows could not appear in both the training and the test sets. The training sets always comprised 75% of the data, and the tests sets comprised the remaining 25% of the data.

Next, we fit a series of ridge regression models to the training data. Specifically, for each pairing of memory, fitness, and demographic variables, we fit a single ridge regression model treating the first variable group as the input features and the second variable group as the target features. For example, one regression model used memory variables to predict fitness variables, and another regression model used fitness variables to predict demographic variables, and so on. In total we fit six regression models to each training dataset. We then applied the fitted models to the held-out test dataset and computed the root mean squared deviation (RMSD) between the predicted and observed values in the target features of the test dataset. We also examined the regression weights assigned to each input feature. This yielded, for each regression model (across N bootstrap iterations) a distribution of RMSD values and a distribution of weights for each input variable.

We constructed a "null" distribution by using the same procedure as above, but where the columns in the test datasets were randomly permuted with each iteration (thereby breaking any meaningful predictive information between the training and test data). We assessed the statistical significance (p-values) of the observed RMSD values by computing the proportions of null RMSD values that were less than the observed value. We also assessed the significance of the observed regression weights using t-tests to compare the means of the observed versus null distributions of weights.

Reverse correlation analyses

We sought to characterize potential associations between the history of participants' fitness-related activities leading up to the time they participated in a memory task and their performance on the given task. For each fitness-related variable, we constructed a timeseries matrix whose rows corresponded to timepoints (sampled once per day) leading up to the day the participant accepted the HIT for our study, and whose columns corresponded to different participants. These matrices often contained missing entries, since different participants' Fitbit devices tracked fitness-related activities differently. For example, participants whose Fitbit devices lacked heart rate sensors would have missing entries for any heart rate-related variables. Or, if a given participant neglected to wear their fitness tracker on a particular day, the column corresponding to that participant would have missing entries for that day.

In addition to this set of matrices storing timeseries data for each fitness-related variable, we also constructed a memory performance matrix, M, whose rows corresponded to different memory-related variables, and whose columns corresponded to different participants. For example, one row of the memory performance matrix reflected the average proportion of words (across lists) that each participant remembered during the immediate free recall test, and so on.

Given a fitness timeseries matrix, F, we computed the weighted average and weighted standard error of the mean of each row of F, where the weights were given by a particular memory-related variable (row of M). For example, if F contained participants' daily step counts, we could use any row of M to compute a weighted average across any participants who contributed step count data on each day. Choosing a row of M that corresponded to participants' performance on the naturalistic recall task would mean that participants who performed better on the naturalistic recall task would contribute more to the weighted average timeseries of daily step counts. Specifically, for each row, t, of F, we computed the weighted average (across the S participants) as:

$$\bar{f}(t) = \sum_{s=1}^{S} \dot{m}(s) F(t, s),$$

where $\dot{m}$ denotes the normalized min-max scaling of m (the row of M corresponding to the chosen memory-related variable):

$$\dot{m} = \frac{m}{\sum_{s=1}^{S} \hat{m}(s)},$$

where

$$\hat{m} = \frac{m - \min(m)}{\max(m) - \min(m)}$$

We computed the weighted standard error of the mean as:

$$\mathrm{SEM}_m\left(f(t)\right) = \frac{\left|\sum_{s=1}^{S}\left(F(t,s) - \bar{f}(t)\right)\right|}{\sqrt{S}}.$$

When a given row of F was missing data from one or more participants, those participants were excluded from the weighted average for the corresponding timepoint and the weights (across all remaining participants) were re-normalized to sum to 1. The above procedure yielded, for each memory variable, a timeseries of average (and standard error of the mean) fitness tracking values leading up to the day of the experiment.

Acknowledgements

We acknowledge useful discussions with David Bucci, Emily Glasser, Andrew Heusser, Abigail Bartolome, Lorie Loeb, Lucy Owen, and Kirsten Ziman. Our work was supported in part by the Dartmouth Young Minds and Brains initiative. The content is solely the responsibility of the authors and does not necessarily represent the official views of our supporting organizations. This paper is dedicated to the memory of David Bucci, who helped to inspire the theoretical foundations of this work. Dave served as a mentor and colleague on the project prior to his passing.

Data and code availability

All analysis code and data used in the present manuscript may be found here.

Author contributions

Concept: J.R.M. Experiment implementation and data collection: G.M.N. Analyses: G.M.N., E.C., P.C.F., and J.R.M. Writing: J.R.M.

Competing interests

The authors declare no competing interests.

References

Bassey, E. J. and Ramsdale, S. J. (1994). Increase in femoral bone density in young women following high-impact exercise. *Osteoporosis International*, 4:72–75.

Basso, J. C. and Suzuki, W. A. (2017). The effects of acute exercise on mood, cognition, neurophysiology, and neurochemical pathways: a review. *Brain Plasticity*, 2(2):127–152.

Brisswalter, J., Collardeau, M., and René, A. (2002). Effects of acute physical exercise characteristics on cognitive performance. *Sports Medicine*, 32:555–566.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society*, 22(2):302–306.

Callaghan, P. (2004). Exercise: a neglected intervention in mental health care? *Psychiatric and Mental Health Nursing*, 11(4):476–483.

Chang, Y. K., Labban, J. D., Gapin, J. I., and Etnier, J. L. (2012). The effects of acute exercise on cognitive performance: a meta-analysis. *Brain Research*, 1453:87–101.

Chilibeck, P. D., Sale, D. G., and Webber, C. E. (2012). Exercise and bone mineral density. *Sports Medicine*, 19:103–122.

Crane, J. D., MacNeil, L. G., and Tarnopolsky, M. A. (2013). Long-term aerobic exercise is associated with greater muscle strength throughout the life span. *The Journals of Gerontology: Series A*, 68(6):631–638.

Deslandes, A., Moraes, H., Ferreira, C., Veiga, H., Silveira, H., Mouta, R., Pompeu, F. A. M. S., Coutinho, E. S. F., and Laks, J. (2009). Exercise and mental health: many reasons to move. *Neuropsychobiology*, 59:191–198.

Etnier, J. L., Nowell, P. M., Landers, D. M., and Sibley, B. A. (2006). A meta-regression to examine the relationship between aerobic fitness and cognitive performance. *Brain Research: Brain Research Reviews*, 52(1):119–130.

Knuttgen, H. G. (2007). Strength training and aerobic exercise: comparison and contrast. *Journal of Strength and Conditioning Research*, 21(3):973–978.

Layne, J. E. and Nelson, M. E. (1999). The effects of progressive resistance training on bone density: a review. *Medicine and Science in Sports and Exercise*, 31(1):25–30.

Lazovic-Popovic, B., Zlatkovic-Svenda, M., Durmic, T., Djelic, M., Saranovic, D., and Zugic, V. (2016). Superior lung capacity in swimmers: some questions, more answers! *Revista Portuguesa de Pneumologia*, 22(3):151–156.

Lindh, M. (1979). Increase of muscle strength from isometric quadriceps exercises at different knee angles. *Scandinavian Journal of Rehabilitation Medicine*, 11(1):33–36.

Maiorana, A., O'Driscoll, G., Cheetham, C., Collis, J., Goodman, C., Rankin, S., Taylor, R., and Green, D. (2000). Combined aerobic and resistance exercise training improves functional capacity and strength in CHF. *Journal of Applied Physiology*, 88(1565–1570).

Mikkelsen, K., Stojanovska, L., Polenakovic, M., Bosevski, M., and Apostolopoulos, V. (2017). Exercise and mental health. *Maturitas*, 106:48–56.

Morton, J. P., Kayani, A. C., McArdle, A., and Drust, B. (2009). The exercise-induced stress response of skeletal muscle, with specific emphasis on humans. *Sports Medicine*, 39:643–662.

Palacios-Filardo, J., Udakis, M., Brown, G. A., Tehan, B. G., Congreve, M. S., Nathan, P. J., Brown, A. J. H., and Mellor, J. R. (2021). Acetylcholine prioritises direct synaptic inputs from entorhinal cortex to CA1 by differential modulation of feedforward inhibitory circuits. *Nature Communications*, 12(5475):doi.org/10.1038/s41467–021–25280–5.

Paluska, S. A. and Schwenk, T. L. (2000). Physical activity and mental health. *Sports Medicine*, 29(3):167–180.

Pollock, M. L., Franklin, B. A., Balady, G. J., Chaltman, B. L., Fleg, J. L., Fletcher, B., Limacher, M., na, I. L. P., Stein, R. A., Williams, M., and Bazzarre, T. (2000). Resistance exercise in individuals with and without cardiovascular disease. *Circulation*, 101:828–833.

Raglin, J. S. (1990). Exercise and mental health. *Sports Medicine*, 9:323–329.

Rogers, M. A. and Evans, W. J. (1993). Changes in skeletal muscle with aging: effects of exercise training. *Exercise and Sport Sciences Reviews*, 21:65–102.

Roman, M. A., Rossiter, H. B., and Casaburi, R. (2016). Exercise, ageing and the lung. *European Respiratory Journal*, 48:1471–1486.

Schiaffino, S., Dyar, K. A., Ciciliot, S., Blaauw, B., and Sandri, M. (2013). Mechanisms regulating skeletal muscle growth and atrophy. *The febs Journal*, 280(17):4294–4314.

Shoemaker, J. K., Halliwill, J. R., Hughson, R. L., and Joyner, M. J. (1997). Contributions of acetylcholine and nitric oxide to forearm blood flow at exercise onset and recovery. *Vascular Physiology*, 273(5):2388–2395.

Taylor, C. B., Sallis, J. F., and Needle, R. (1985). The relation of physical activity and exercise to mental health. *Public Health Reports*, 100(2):195–202.

Wilmore, J. H. and Knuttgen, H. G. (2003). Aerobic exercise and endurance. *The Physician and Sportsmedicine*, 31(5):45–51.

Ziman, K., Heusser, A. C., Fitzpatrick, P. C., Field, C. E., and Manning, J. R. (2018). Is automatic speech-to-text transcription ready for use in psychological experiments? *Behavior Research Methods*, 50:2597–2605.